基因组学数据分析 第一次作业

```
Environment:
    R version 3.5.2 (2018-12-20)
    Platform: x86_64-pc-linux-gnu (64-bit)
    Running under: Debian GNU/Linux 10 (buster)
    Coding: UTF-8
```

College 数据集分析

1. 读入数据

```
college <- read.csv("./College.csv")
```

2. 观察数据

使用`class()`判断数据类型，使用`dim()`查看大小

```
class(college)
```

```
[1] "data.frame"
```

```
dim(college)
```

```
[1] 777  19
```

为 777×19 的 `data.frame` 类

3. 重置行名

原本的行列

```
head(college)
```

```
                             X Private Apps Accept Enroll Top10per
c Top25perc
1 Abilene Christian University     Yes 1660   1232    721        2
3        52
2           Adelphi University     Yes 2186   1924    512        1
6        29
3                Adrian College     Yes 1428   1097    336        2
2        50
```

```
4          Agnes Scott College     Yes  417    349    137        6
0        89
5    Alaska Pacific University     Yes  193    146     55        1
6        44
6            Albertson College     Yes  587    479    158        3
8        62
  F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD T
erminal
1        2885         537     7440       3300   450     2200  70
78
2        2683        1227    12280       6450   750     1500  29
30
3        1036          99    11250       3750   400     1165  53
66
4         510          63    12960       5450   450      875  92
97
5         249         869     7560       4120   800     1500  76
72
6         678          41    13500       3335   500      675  67
73
  S.F.Ratio perc.alumni Expend Grad.Rate
1      18.1          12   7041        60
2      12.2          16  10527        56
3      12.9          30   8735        54
4       7.7          37  19016        59
5      11.9           2  10922        15
6       9.4          11   9727        55
```

重置后行列

```
rownames(college) <- college[, 1]
college <- college[, -1]
head(college)
```

```
                           Private Apps Accept Enroll Top10perc
 Top25perc
Abilene Christian University     Yes 1660   1232    721        23
52
Adelphi University               Yes 2186   1924    512        16
29
Adrian College                   Yes 1428   1097    336        22
50
Agnes Scott College              Yes  417    349    137        60
89
Alaska Pacific University        Yes  193    146     55        16
44
Albertson College                Yes  587    479    158        38
62
                           F.Undergrad P.Undergrad Outstate Roo
m.Board Books
Abilene Christian University        2885         537     7440
3300   450
Adelphi University                  2683        1227    12280
6450   750
Adrian College                      1036          99    11250
3750   400
```

```
Agnes Scott College                 510          63    12960
5450   450
Alaska Pacific University           249         869     7560
4120   800
Albertson College                   678          41    13500
3335   500
                            Personal PhD Terminal S.F.Ratio perc.
alumni Expend
Abilene Christian University     2200  70       78      18.1
12   7041
Adelphi University               1500  29       30      12.2
16  10527
Adrian College                   1165  53       66      12.9
30   8735
Agnes Scott College               875  92       97       7.7
37  19016
Alaska Pacific University        1500  76       72      11.9
2  10922
Albertson College                 675  67       73       9.4
11   9727
                            Grad.Rate
Abilene Christian University        60
Adelphi University                  56
Adrian College                      54
Agnes Scott College                 59
Alaska Pacific University           15
Albertson College                   55
```

```
college$Private <- as.numeric(as.factor(college$Private))
```

4. 计算Apps和Accept的平均值与相关系数

   - 计算平均值

```
apps.mean <- mean(college$Apps)
apps.mean
```

```
[1] 3001.638
```

```
accept.mean <- mean(college$Accept)
accept.mean
```

```
[1] 2018.804
```

   - 计算相关系数

```
lm.res <- lm(Accept ~ Apps, college)
summary(lm.res)
```

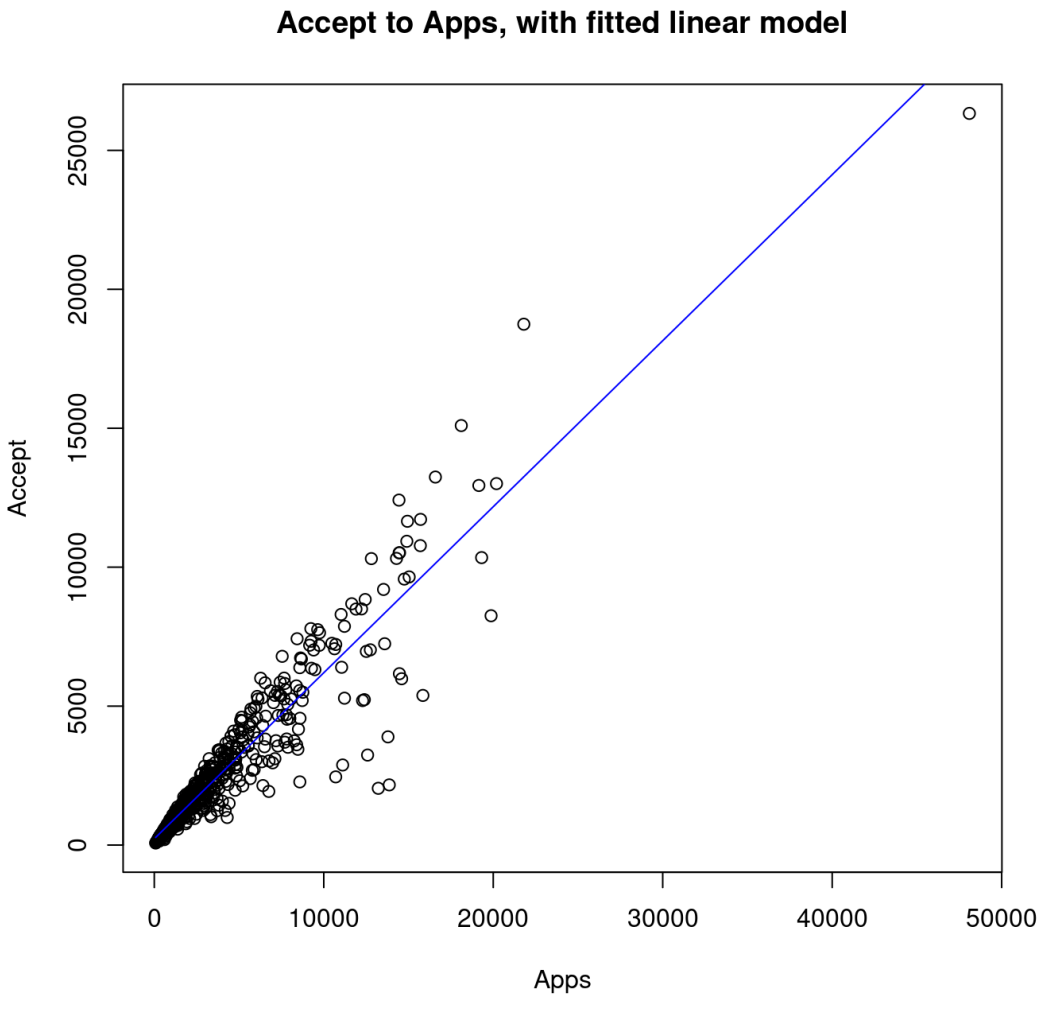
```
Call:
```

```
lm(formula = Accept ~ Apps, data = college)

Residuals:
    Min      1Q  Median      3Q     Max
-6344.8  -154.2   -35.2   184.7  5490.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.253e+02  3.692e+01   6.101 1.66e-09 ***
Apps        5.975e-01  7.542e-03  79.226  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 813.1 on 775 degrees of freedom
Multiple R-squared:  0.8901,    Adjusted R-squared:   0.89
F-statistic:  6277 on 1 and 775 DF,  p-value: < 2.2e-16
```

可以看到R^2统计量为0.8901，而在简单线性模型中，相关系数的平方有$r^2 = R^2$的关系，因此不难得到$r = \sqrt{R^2} = 0.943$

- 绘制拟合结果

```
plot(college$Apps, college$Accept, xlab = "Apps", ylab = "Accept",
lines(college$Apps, fitted(lm.res), col = "blue")
```



5. `summary()`函数的使用

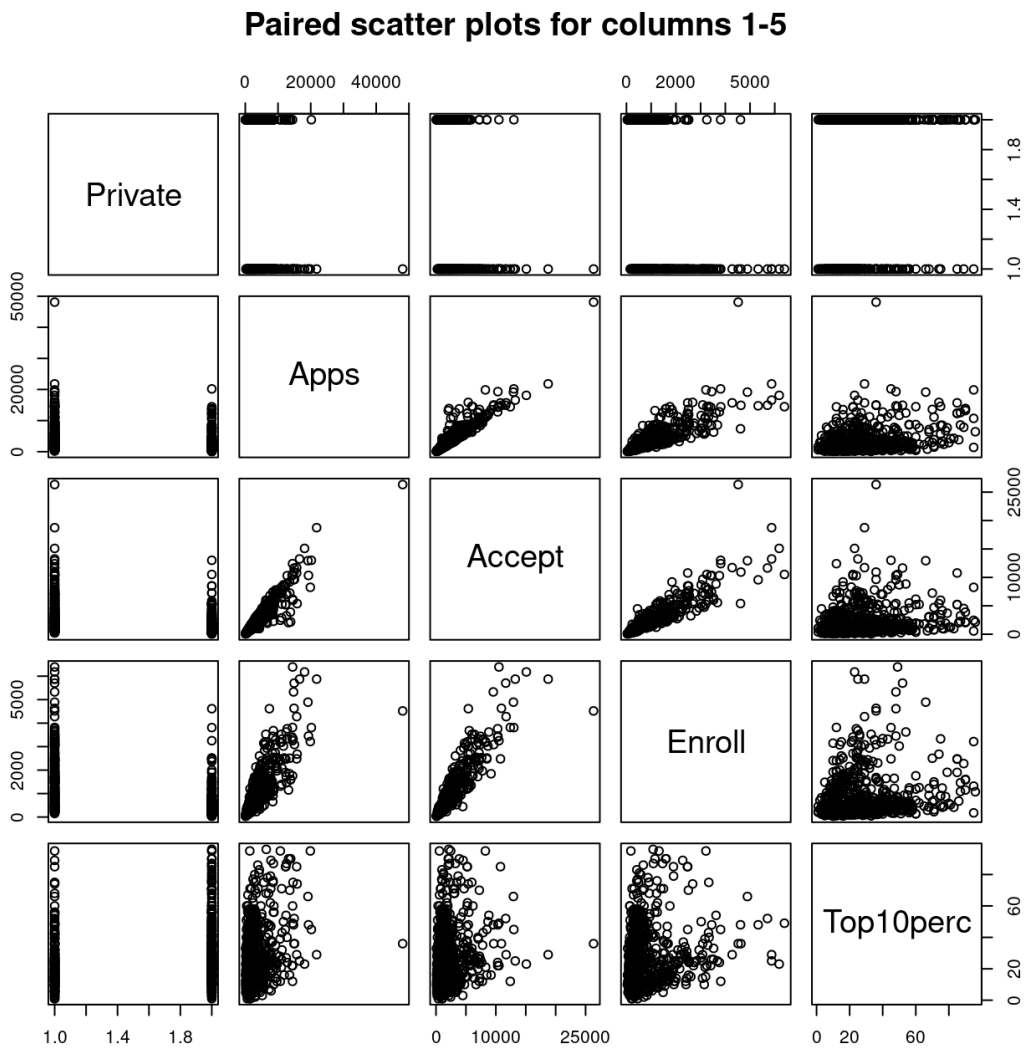
```
summary(college[, 1:3])
```

```
    Private           Apps           Accept
Min.   :1.000   Min.   :   81   Min.   :   72
1st Qu.:1.000   1st Qu.:  776   1st Qu.:  604
Median :2.000   Median : 1558   Median : 1110
Mean   :1.727   Mean   : 3002   Mean   : 2019
3rd Qu.:2.000   3rd Qu.: 3624   3rd Qu.: 2424
Max.   :2.000   Max.   :48094   Max.   :26330
```

其中`1st Qu.`为第一四分位点，表示约有25%的数据在此值以下；`3rd Qu.`为第三四分位点，表示约有75%的数据在此值以下

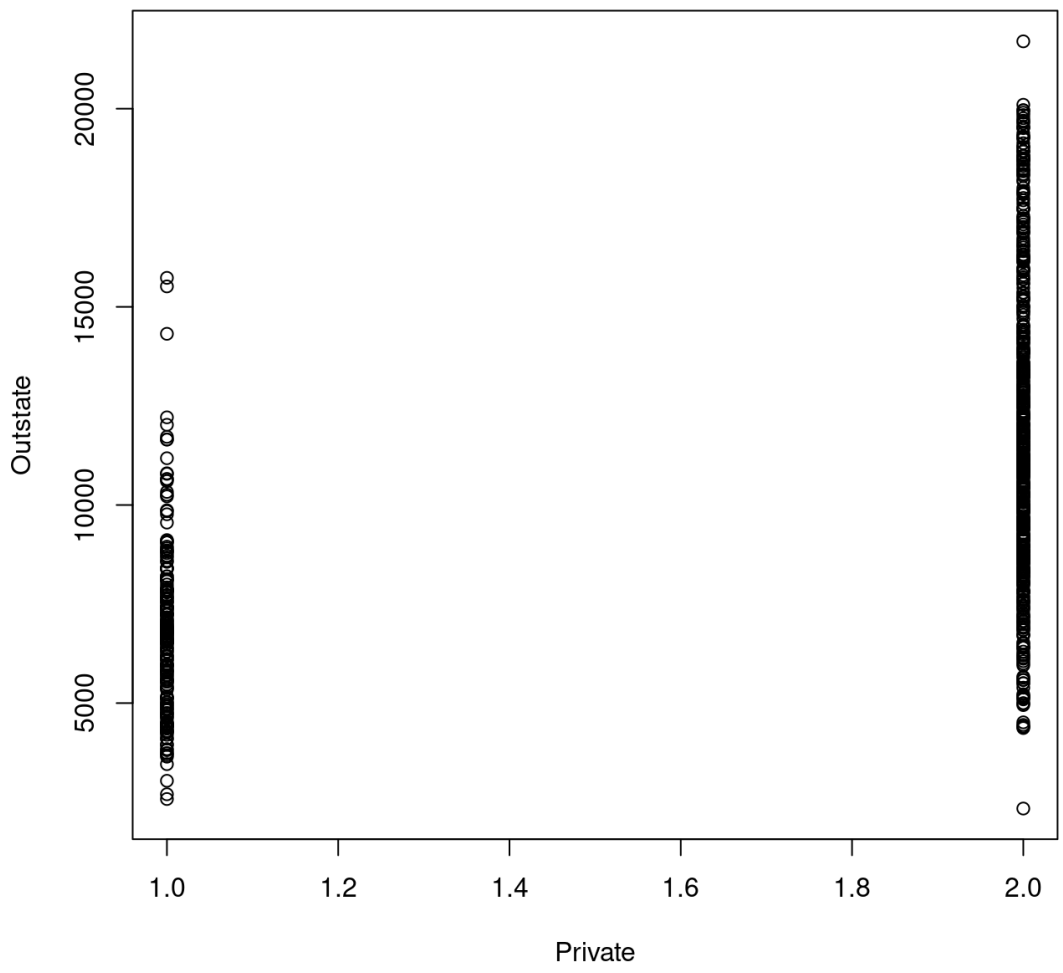
6. `pairs()`函数作图

```
pairs(college[, 1:5], main = "Paired scatter plots for columns 1-5")
```

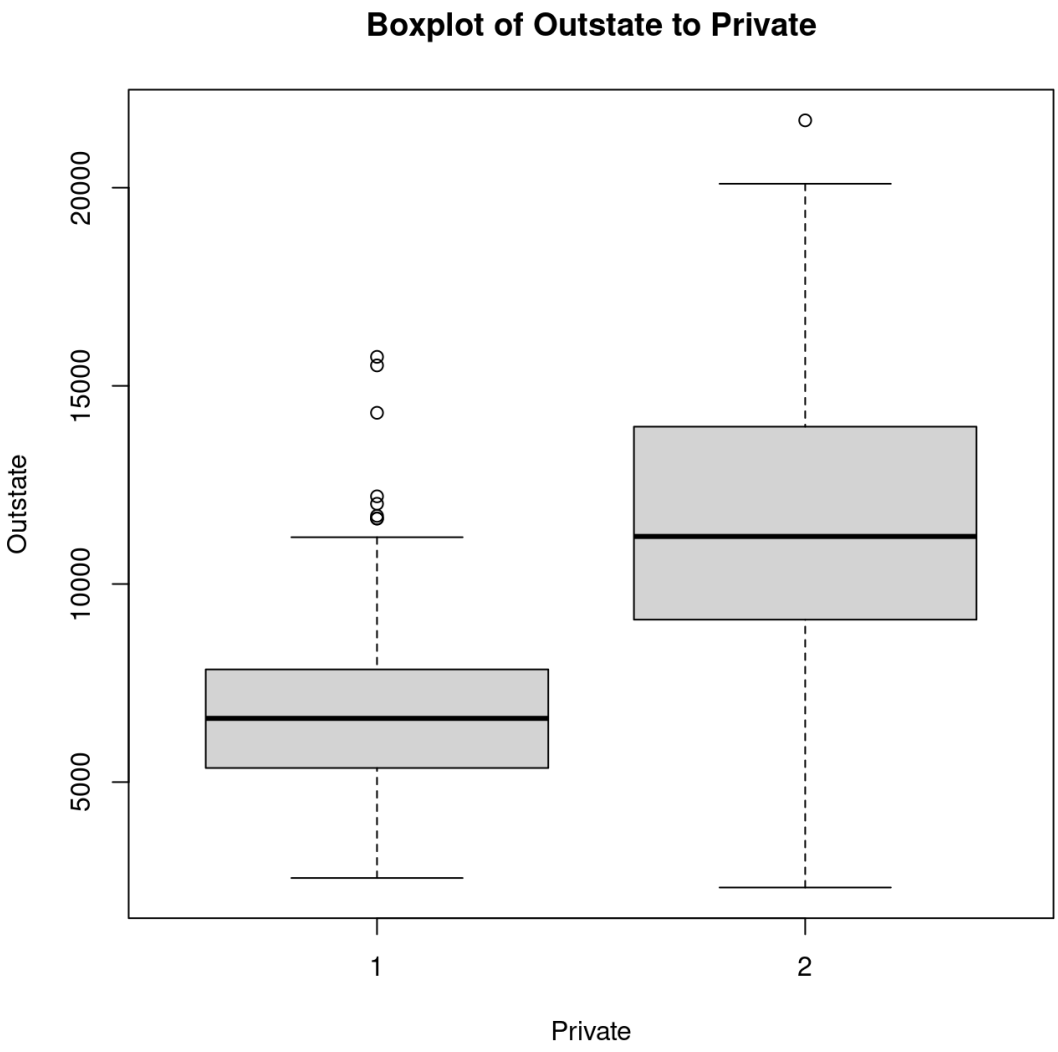


7. `plot()`与`boxplot()`作图

```
plot(college$Private, college$Outstate, ylab = "Outstate", xlab = "Pri
```




```
boxplot(college$Outstate ~ college$Private, ylab = "Outstate", xlab =
```



> `boxplot()`函数第一项参数为`formula`，而不接受分开的X与Y
>
> `plot()`函数的默认作图方式也为箱线图，与`boxplot()`一致

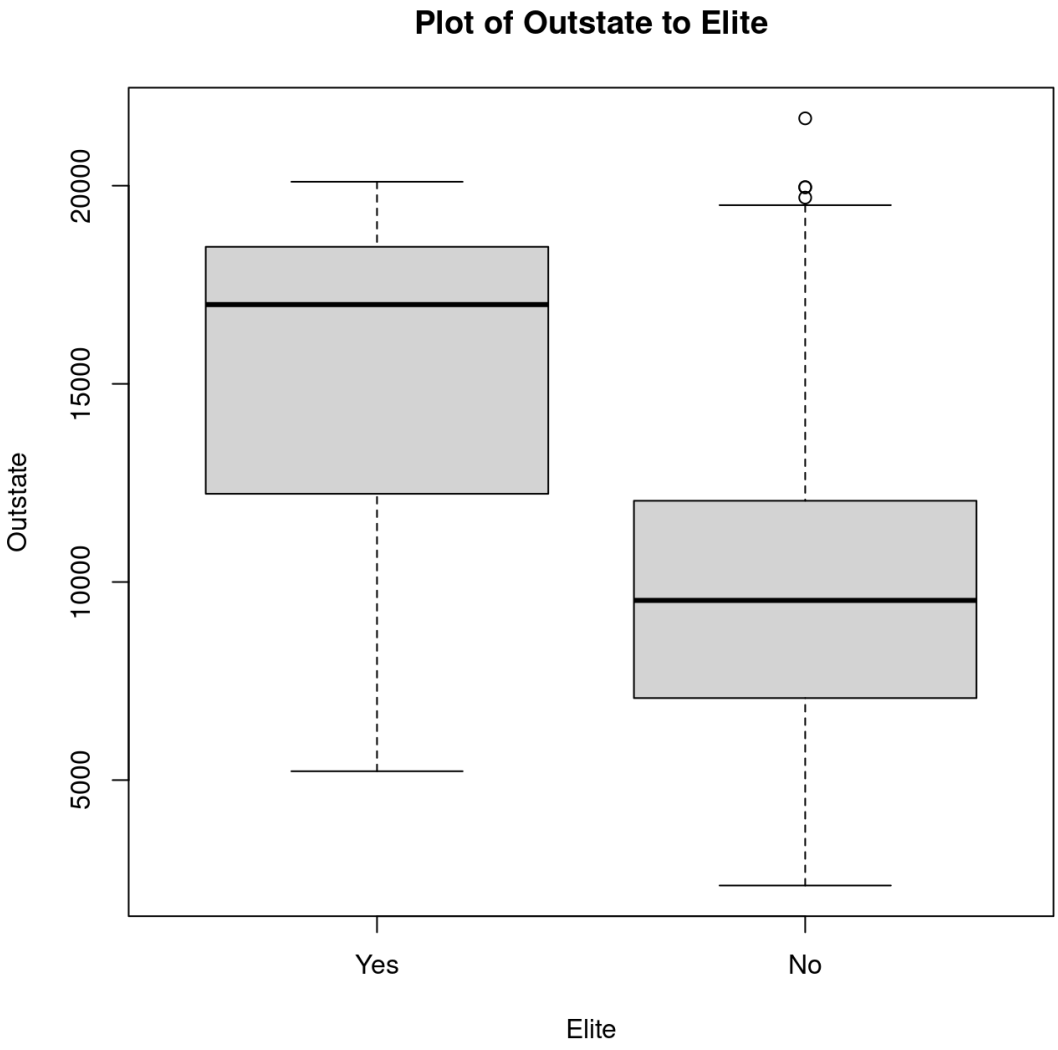
8. 新增`Elite`列并作图

```
college["Elite"] <- ifelse(college$Top10perc >= 50, "Yes", "No")
head(college)
```

```
                           Private Apps Accept Enroll Top10perc
 Top25perc
Abilene Christian University       2 1660   1232    721        23
52
Adelphi University                 2 2186   1924    512        16
29
Adrian College                     2 1428   1097    336        22
50
Agnes Scott College                2  417    349    137        60
89
Alaska Pacific University          2  193    146     55        16
44
Albertson College                  2  587    479    158        38
62

                           F.Undergrad P.Undergrad Outstate Roo
```

```
m.Board Books
Abilene Christian University        2885         537     7440
3300   450
Adelphi University                  2683        1227    12280
6450   750
Adrian College                      1036          99    11250
3750   400
Agnes Scott College                  510          63    12960
5450   450
Alaska Pacific University            249         869     7560
4120   800
Albertson College                    678          41    13500
3335   500
                            Personal PhD Terminal S.F.Ratio perc.
alumni Expend
Abilene Christian University     2200  70       78      18.1
12   7041
Adelphi University               1500  29       30      12.2
16  10527
Adrian College                   1165  53       66      12.9
30   8735
Agnes Scott College               875  92       97       7.7
37  19016
Alaska Pacific University        1500  76       72      11.9
2  10922
Albertson College                 675  67       73       9.4
11   9727
                            Grad.Rate Elite
Abilene Christian University        60    No
Adelphi University                  56    No
Adrian College                      54    No
Agnes Scott College                 59   Yes
Alaska Pacific University           15    No
Albertson College                   55    No
```

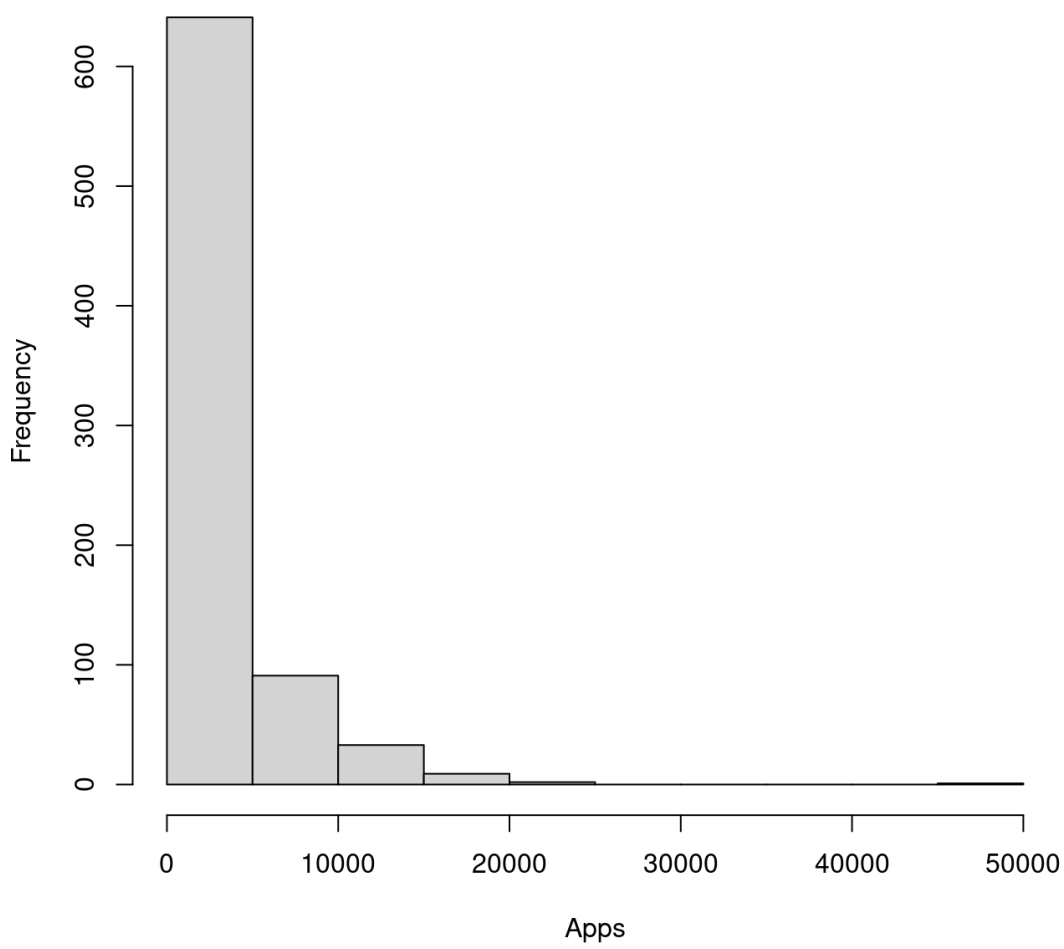
```
plot(factor(college$Elite, levels = c("Yes", "No")), college$Outstate,
```



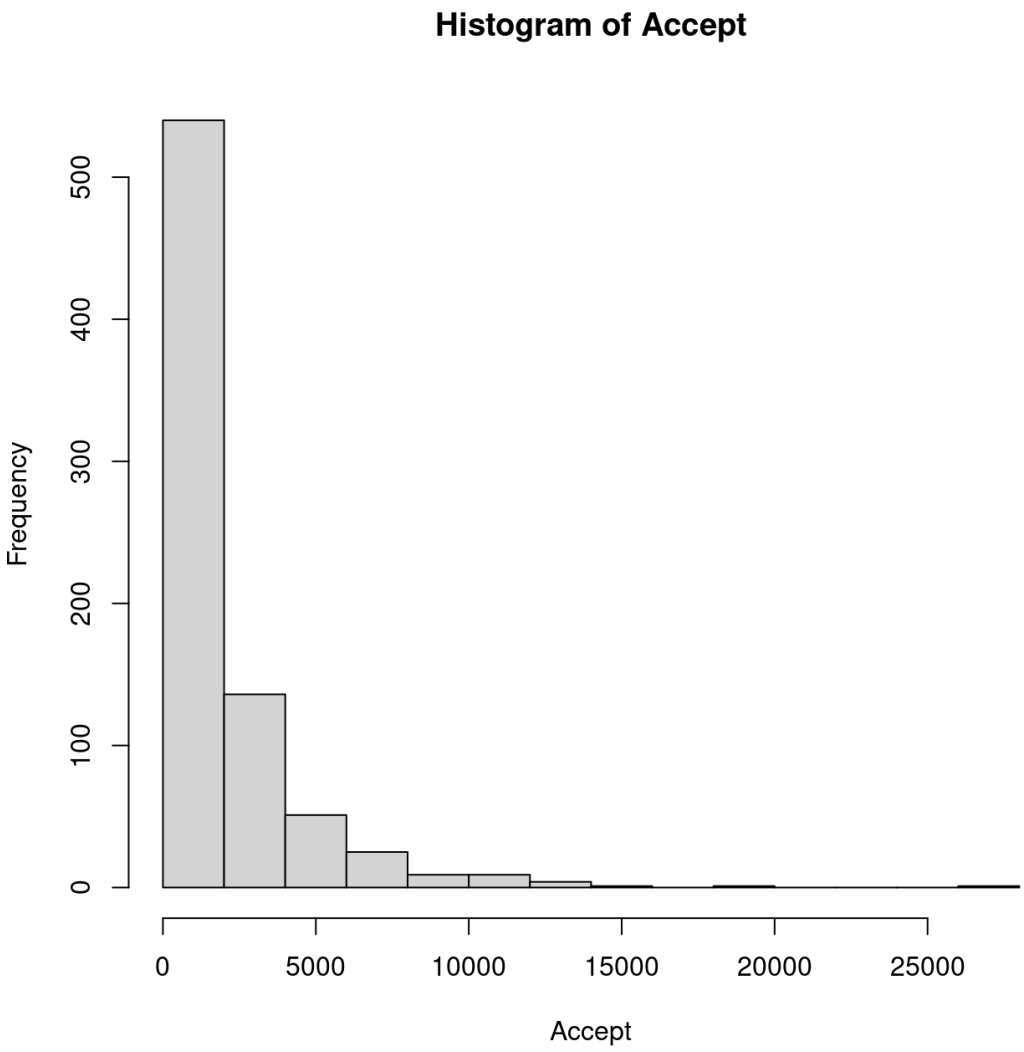
> 这里使用了`factor()`函数内的`levels`参数，将`factor`类型的默认排序由字典序变为了自定义的顺序

9. `hist()`函数作图

```
hist(college$Apps, xlab = "Apps", main = "Histogram of Apps")
```




```
hist(college$Accept, xlab = "Accept", main = "Histogram of Accept")
```



> 此处可以不传入`main`参数，`hist()`对此参数有默认取值

10. 数据写入本地

```
write.csv(college, file = "./College.new.csv")
```

Auto 数据集分析

11. 读入数据

    此数据集中存在缺失值，不方便处理，故将整行刨去

```
auto <- read.csv("./Auto.csv", na.string = "?")
auto <- na.omit(auto)
```

12. 简单线性回归

```
lm.res <- lm(mpg ~ horsepower, auto)
summary(lm.res)
```

```
Call:
lm(formula = mpg ~ horsepower, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
confint(lm.res)
```

```
                2.5 %     97.5 %
(Intercept) 38.525212 41.3465103
horsepower  -0.170517 -0.1451725
```

- **预测变量和响应变量之间有关系吗**

  有关系（$R^2 = 0.6059$）

- **关系有多强**

  从$R^2 = 0.6059$来看，仅有不到$\frac{2}{3}$的响应变量中的差异能够被预测变量所解释，因此关系并不强

- **是正相关还是负相关**

  从`horsepower`的系数为$-0.157845 < 0$来看，为负相关

- **当`horsepower`为98时，`mpg`的预测值为多少，相应的95%置信区间与预测区间分别为多少**

  ```
  predict(lm.res, data.frame(horsepower = 98), interval = "confidenc
  ```

  ```
         fit      lwr      upr
  1 24.46708 23.97308 24.96108
  ```

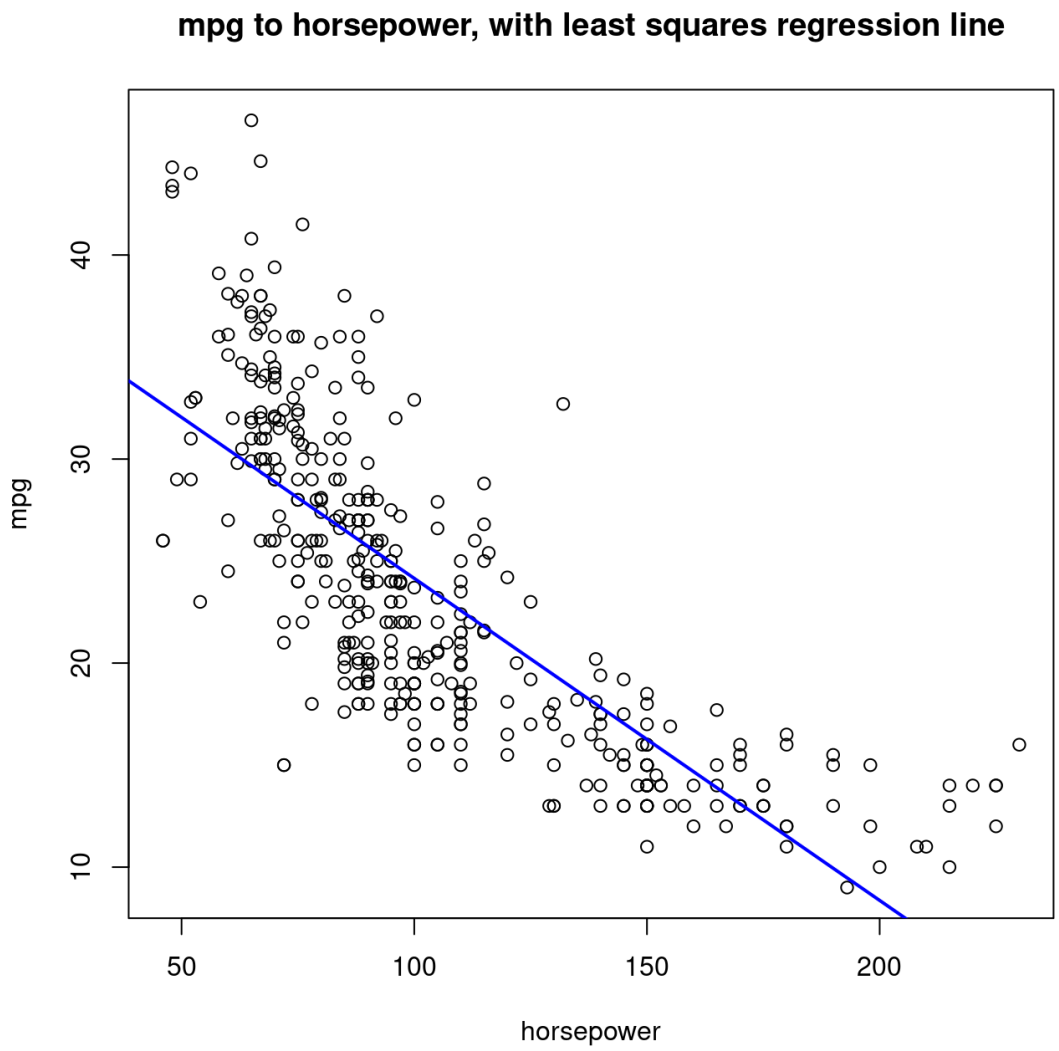
  ```
  predict(lm.res, data.frame(horsepower = 98), interval = "predictio
  ```

  ```
         fit     lwr      upr
  1 24.46708 14.8094 34.12476
  ```

故 `mpg` 预测值为24.46708，95%置信区间为$(23.97308, 24.96108)$，预测区间为$(14.8094, 34.12476)$

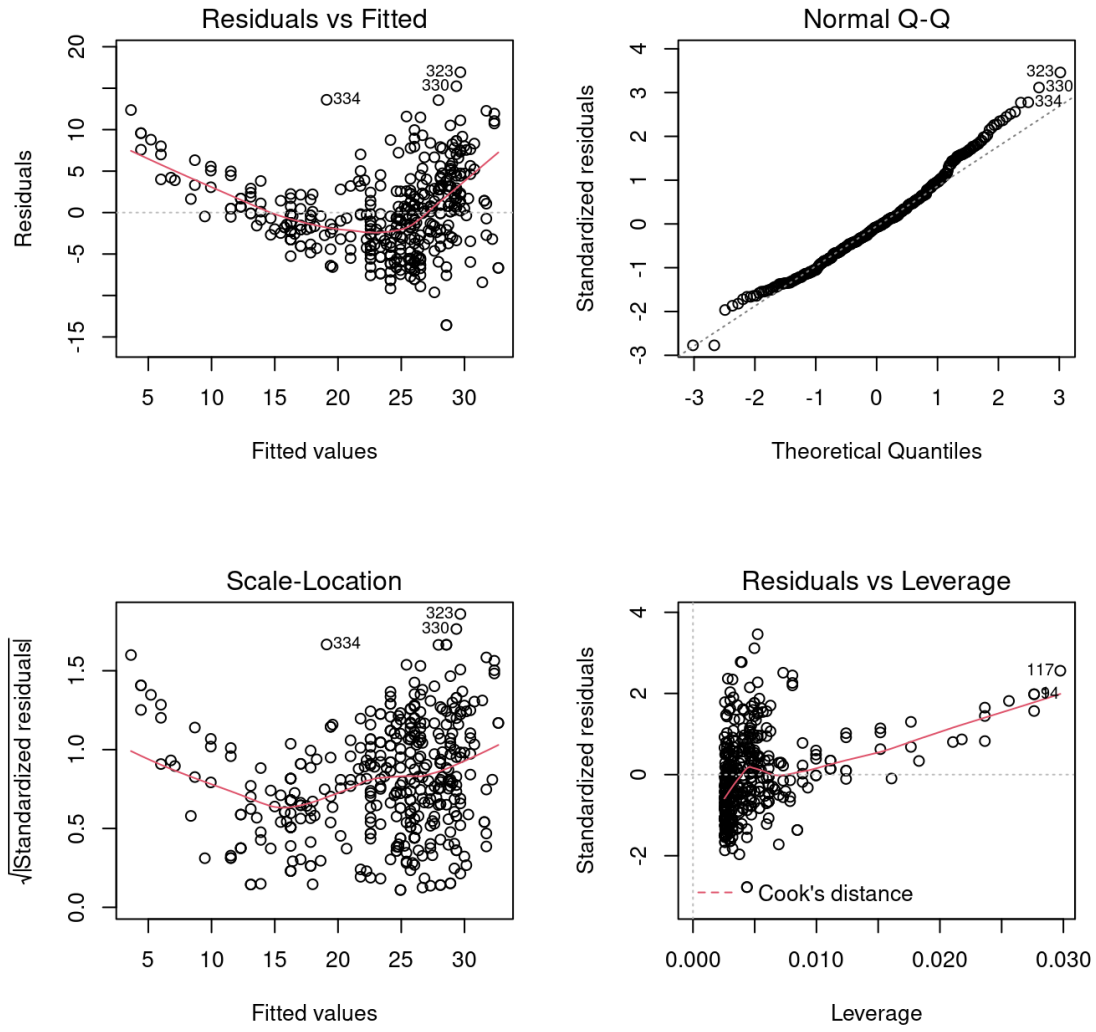
13. 绘制关系图与最小二乘回归线

```
plot(auto$horsepower, auto$mpg, xlab = "horsepower", ylab = "mpg", mai
abline(lm.res, lwd = 2, col = "blue")
```



14. 最小二乘回归拟合诊断图

```
par(mfrow = c(2, 2))
plot(lm.res)
```



从图中不难看出，只有靠近中段百分位的拟合程度较好。观察杠杆值图，不难发现杠杆值较小的点随在一较大的残差波动范围内，但总体残差较小，比较集中；杠杆值大致处于中间的数据点残差相对来说较小；杠杆值大的数据点又明显有较大残差；反映在结果上就是两端的数据点和中间腹部的数据点都拟合较差，出现了相反方向的偏离预测模型。